The American National Corpus First Release

Nancy Ide and Keith Suderman
Department of Computer Science, Vassar College, Poughkeepsie, NY 12604-0520 USA
ide@cs.vassar.edu, suderman@cs.vassar.edu

Abstract

The First Release of the American National Corpus (ANC) was made available in mid-fall, 2003. The data includes approximately 11 million words of American English, including written and spoken data and a variety of text types annotated for part of speech and lemma. The corpus is provided in XML format conformant to the XML Corpus Encoding Standard (XCES) (http://www.xml-ces.org), and is distributed in both a stand-off version (where annotation is in an XML document separate from the primary texts) and a merged version (where annotation is included in-line in the texts). The merged version includes annotation for part of speech and lemma produced by the Biber tagger; in stand-off annotation, in addition to the Biber tagging, morpho-syntactic annotations of the data are provided using the CLAWS 5 and 7 tagsets as well as several other tagsets.

Introduction

The American National Corpus (ANC) project is building a large-scale, representative corpus of American English, comparable in size and distribution of genres to the British National Corpus but also including a potentially substantial additional component comprised of materials drawn from varied sources representing a wide range of modern English usage. All materials in the corpus are drawn from written materials and transcripts of spoken data produced from 1990 onward. As such, the American National Corpus will become the definitive record of contemporary American English, invaluable as a resource for the creation of dictionaries, thesauri, and materials for teaching English, as well as the development of software for search and access.

The ANC's first release of 11 million words of data, made available though the Linguistic Data Consortium in October 2003, has been met with enthusiasm by the computational linguistics and linguistics communities. The ANC First Release is available for research and education for a nominal licensing fee from the LDC; commercial users can obtain the corpus and gain rights to use it in commercial products by joining the ANC Consortium. The ANC Consortium and project are described at http://AmericanNationalCorpus.org. For additional information, see Ide & Macleod, 2001; Ide, *et al.*, 2002; and Reppen & Ide, in press.

This paper describes the content and format of the ANC First Release data and outlines the future activities of the ANC project.

ANC First Release Data

Table 1 summarizes the contents of the ANC First Release, consisting of over 3 million words of spoken data and over 8 million words of written data. The first release contains texts first received by the ANC, and is therefore not balanced for genre, and structural markup and part of speech annotation were produced automatically and not hand-validated. Headers are minimal, although they contain fairly complete information concerning *domain, subdomain, subject, audience,* and *medium*. One of the aims of releasing the First Release sub-corpus is to get feedback from the community about its structure and annotation, so that modifications can be made, if necessary, for the final release of the full 100 million word corpus. We therefore invite comments and bug reports from the community of ANC users.

A complete description of the ANC first release is available at

`http://AmericanNationalCorpus.org/FirstRelease`.

ANC Encoding Conventions

The ANC is encoded in XML, conformant to the XML Corpus Encoding Standard (XCES) (Ide, *et al.*, 2000) schemas for primary data and annotations, which are included with the ANC First Release data on the CD distributed by the LDC.

The texts in the corpus are marked to the level of the paragraph, and within paragraphs, for sentence boundaries. Following XCES recommendations, a "stand-off" annotation strategy is followed, meaning that annotations are contained in a separate XML document linked to the original. The associated annotation files identify word (token) boundaries and provide the morpho-syntactic description (part of speech) and lemma for each token in the corpus. Because few processors handle stand-off annotation at this time, a "merged" version of the corpus is also provided, in which each token is explicitly marked with <TOK> tags, and part-of-speech and lemma are given as the values of *msd* and *base* attributes, respectively.

Text Type	Text name	No. of texts	No. of words	Contributor
Spoken	Callhome	24	50,494	LDC
Spoken	Switchboard	2320	3,056,062	LDC
Spoken	Charlotte Narrative	95	117,832	Project MORE
TOTAL TYPES BY POS SPOKEN			**43,864**	
TOTAL WORD TYPES SPOKEN			**22,321**	
TOTAL WORD TOKENS SPOKEN			**3,224,388**	
Written	New York Times	4148	3,207,272	LDC
Written	Berlitz Travel Guides	101	514,021	Langensheidt Publishers
Written	Slate Magazine	4694	4,338,498	Microsoft
Written	Various non-fiction	27	224,037	Oxford University Press
TOTAL TYPES BY POS WRITTEN			**172,192**	
TOTAL WORD TYPES WRITTEN			**96,147**	
TOTAL WORD TOKENS WRITTEN			**8,283,828**	
TOTAL CORPUS TYPES BY POS			**216,056**	
TOTAL CORPUS WORD TYPES			**118,468**	
TOTAL CORPUS WORD TOKENS			**11,508,216**	

Table 1. Contents of the ANC First Release

All primary ANC written documents currently contain sentence boundary markup; however, we realize that sentence markup is a type of linguistic annotation that may vary depending on the particular linguistic theory and/or processing software applied to the data. Therefore, sentence boundary markup will not be included in the primary data in future ANC releases. Spoken data is marked for turn and utterance.

Inter-document linkage to associate stand-off annotations with the primary data uses the W3C XPointer recommendations (DeRose, *et al.*, 2002). The final XPointer recommendations provide means to link down to the level of the element, whereas the ANC data require addressing text within elements. At present, sub-element linkage is accomplished using the string-range mechanism defined in an earlier draft of the XPointer recommendation. The W3C recommends the development of "parenthesized schemes" for sub-element addressing but will not specify a single scheme as the standard. The International Standards Organization (ISO) sub-committee for Language Resource Management is developing a parenthesized scheme for sub-element addressing that should ultimately be adopted as an ISO standard. The ANC will adopt this scheme and provide a script to transduce ANC data to conform to this standard as soon as it becomes available.

Figures 1-3 provide examples of ANC First Release encoding, for both the stand-off and merged versions.

```
<p id="p3">
  <s id="p3s1">
    Ireland has been inhabited since very ancient times, but Irish history really
    begins with the arrival of the Celts around the 6th century b.c.</s>
```

Figure 1. ANC encoding for primary data.

```
<chunk type="sentence" xml:base="#p3s1">
  <tok xlink:href="xpointer(string-range('',0,7))">
     <msd>np++++</msd>
     <base>ireland</base>
  </tok>
  <tok xlink:href="xpointer(string-range('',8,11))">
     <msd>vbz+hvz+aux++</msd>
     <base>have</base>
  </tok>
  <tok xlink:href="xpointer(string-range('',12,16))">
     <msd>vprf+ben+aux+xvbnx+</msd>
     <base>be</base>
  </tok>
  …
```

Figure 2. Stand-off part of speech annotation document for the document fragment in Figure 1.

```
<p id="p3">
  <s id="p3s1">
    <tok msd="np++++" base="ireland">Ireland</tok>
    <tok msd="vbz+hvz+aux++" base="have">has</tok>
    <tok msd="vprf+ben+aux+xvbnx+" base="be">been</tok>
    …
```

Figure 3. "Merged" version of ANC First Release data

Annotations

The default part of speech tagset for the ANC data is the set developed by Biber (1988, 1995), which is included in both the merged and stand-off versions of the Frist Release distributed by the LDC. In addition, Lancaster University in the United Kingdom has generated part-of-speech tags for the ANC First Release conformant to the both the C5[1] and C7 [2] versions of the CLAWS tagset that was used to tag the BNC. We have also generated POS annotations with GATE, which uses a version of the Penn Tags[3], and with the Multext tagger (Gilbert & Armstrong, 1995), which uses the EAGLES tagset for English[4]. These alternative POS annotations each exist as separate stand-off annotation documents linked to the data and distributed via the ANC website. Although for copyright reasons we cannot make the corpus itself freely downloadable from the website without licensing, we can distribute the annotations, as long as the data itself is not included. To access the original data via the links in the annotation documents, the user must obtain the corpus through the LDC and sign the appropriate licenses. A script to create a merged version of the data using any of the annotations provided is also available from the ANC website.

We should note that multiple alternative annotations for the ANC data not only provide annotations suited to different schemes and linguistic theories, but also enable the comparison and merging of these annotations that could lead to methods for disambiguating automatically-produced tags.[5] For example, combining the results of multiple part of speech taggers has been previously shown to be a viable means to produce a more accurate tagging (Brill & Wu, 1998; Tufis, 2000; Sjöbergh, 2003). Also, since part of speech tagsets are designed according to varying criteria (granularity, more or less semantic information, etc.), the availability of a massive corpus annotated with different tagsets can provide information for comparison of linguistic theories.

The American National Corpus project has been awarded a grant from the National Science Foundation to produce a 10 million word "gold standard" corpus, in which all structural tags (e.g. tags marking text divisions, titles, footnotes, lists, plus sub-paragraph markup for abbreviations, names, dates, and items highlighted in the original text) and morpho-syntactic annotation have been hand-validated to assure near 100% accuracy. Hand-corrected morpho-syntactic analysis of the gold standard corpus can be used in conjunction with the output of multiple taggers to increase accuracy of a newly-trained classifier. The gold standard corpus will be available in 2005.

In addition to part of speech annotations, Dekang Lin at the University of Alberta has created a dependency database from the ANC first release data, together with a dependency grammar-based syntactic annotation. This annotation is available as a separate stand-off file linked to the original data and is freely distributed from the ANC website.

Retrieval Software

Because of its XML format, the ANC can be processed to generate concordances, retrieve collocations, etc. using XML-aware corpus processing software such as MonoConc[6] and X/SARA[7], the BNC's XML version of SARA. However, no available software for processing corpora is platform-independent, nor can it be distributed as open source. The ANC project is committed to the production and, to the extent possible, the use of open source, platform-independent, and freely available software. Our goal is to enable users of the corpus—especially users with limited computing expertise—to easily use any tools produced by the project, and/or use the same processing tools (and any supporting resources we develop that work with these tools) to manipulate the corpus and associated data. We are therefore developing a simple query program in Java called AQS (ANC Query System) that uses the open source Berkeley DB 4.1.25 from Sleepycat Software[8] for storing indices that are used to retrieve XML fragments. This design permits later versions of AQS to index and query any corpus that conforms to the XCES schemas.

A beta version of AQS (requiring, again, acquisition of the ANC data from the LDC) is downloadable from the ANC website. Because it is implemented in Java, AQS is platform-independent (requiring only Java 1.4 to run on under any operating system). The beta version generates concordances only, and it is not possible to narrow queries by part of speech, domain, or other features. However, future versions of AQS will be able to perform collocations and filter the results by part of speech, medium, domain, and lemma.

[1] http://www.comp.lancs.ac.uk/ucrel/claws5tags.html

[2] http:// www.comp.lancs.ac.uk/ucrel/claws7tags.html

[3] http://americannationalcorpus.org/FirstRelease/gatetags.txt

[4] ftp://ftp.ilc.pi.cnr.it/pub/eagles/lexicons/elm_en.ps.gz

[5] Note that the segment boundaries for words (tokens) are contained in the links of the stand-off annotation document, and therefore segmentation differences among taggers and syntactic analyzers are not a problem

[6] http://www.ruf.rice.edu/~barlow/mono.html

[7] http://www.oucs.ox.ac.uk/rts/xara/

[8] http://www.sleepycat.com/

The ANC, Present and Future

We are currently processing texts for a second release of ANC data, which will be roughly the same size as the First Release and appear later in 2004. Among the data in the processing pipeline are non-fiction books contributed by Cambridge University Press and Pearson Education, several millions of words of public domain U.S. government documents taken from the web, Verbatim Magazine, "ephemera" consisting of flyers and pamphlets, biomedical articles, and several additional spoken corpora including a subset of the Michigan Corpus of Academic Spoken English (MICASE).[9] We have also acquired logs of a "Buffy the Vampire Slayer" discussion group and several novels contributed or placed in the public domain by independent authors.

We are exploring a number of avenues by which to continue to acquire texts for the ANC, and in particular substantially more fiction than we currently have on hand. One strategy was to set up a web site accessible from the ANC home page through which American authors can contribute texts produced after 1989 for inclusion in the ANC—with the reward of knowing that their language might have an impact on dictionaries and other linguistic resources developed in the future. We are also acquiring data from "bloggers", much of which will likely be included in the portion of the ANC outside the 100 million word core, but which provides perhaps one of the more interesting samples of American English as it is currently used on a day-to-day basis.

A Linguistic Infrastructure for American English

As the ANC has developed, its potential to provide not only a corpus, but also a wide variety of linguistic materials and data derived from it, has become apparent. We now envision the creation of a comprehensive "linguistic infrastructure" for American English based on the ANC, including not only basic resources like frequency wordlists, bigrams and trigrams for both tokens and POS, etc., but also multiple annotations for part of speech and syntax using different tagsets, annotation for co-reference and named entities, annotation by semantic categories that link the ANC data to categories in WordNet and FrameNet, categories and ontologies describing and linking linguistic information in the ANC, and comparative data for British and American English such as syntactic and lexical variants, etc. No such repository of linguistic information of this kind and scope exists for corpora in any language. By making resources of this kind available, the ANC project will provide an invaluable and much needed resource for linguistics, computational linguistics, corpus linguistics, and lexicographic research, including cross-linguistic studies, language acquisition, and English language education. In addition, given our commitment to using state-of-the-art representation models for linguistic data in the ANC project, these materials will comprise an instantiation of such models for linguistic data that can be integrated into the developing framework of the semantic web.

It is not possible for the ANC project to produce all of the resources envisaged as a part of a linguistic infrastructure. We will therefore rely on the contributions and expertise of our colleagues, especially those within the computational linguistics community, to provide software and/or results from processing the ANC data for free distribution.

[9] ACL has contributed one million words of research articles from proceedings since 1990, but the data is likely unusable because of problems retrieving text from PDF files in double-column format.

Conclusion

The American National Corpus is a major resource for linguistic research, as well as lexicography, computational linguistics research, corpus linguistic research, and a resource for the development of English language teaching materials. Because we expect the corpus to be continually expanded in the future, it will provide not only the definitive record of American English at the turn of the millennium, but also a record of the development of American English in the 21st century. Enhancement of the corpus with additional annotation and derived information will substantially increase its value as a resource for research and education.

Acknowledgements

This work is supported by the American National Corpus Consortium. Our thanks also go to the Linguistic Data Consortium for their support of the ANC Project.

References

Biber, D. (1988). Variation across speech and writing. New York: Cambridge University Press.

Biber, D. (1995). Dimension of register variation. New York: Cambridge University Press.

Brill, E. & Wu, J. (1998). Classifier Combination for Improved Lexical Disambiguation. Proceedings of COLING-ACL'98, 191-195.

Cunningham. H. (2002) GATE, A General Architecture for Text Engineering. Computers and the Humanities, 36(2), 223—254.

DeRose, S., Daniel R. Jr., Maler, E. (1999) XML Pointer Language (XPointer), http://www.w3.org/TR/WD-xptr.

Gilbert, R. & Amstrong,, S. (1995). Tagging tool. MULTEXT Deliverable 2.4.1.

Ide, N., Bonhomme, P., & Romary, L. (2000). XCES: An XML-based Encoding Standard for Linguistic Corpora. Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association, 825-30.

Ide, N., Macleod, C. (2001). The American National Corpus: A Standardized Resource of American English. Proceedings of Corpus Linguistics 2001, Lancaster UK.

Ide, N., Reppen, R., & Suderman, K. (2002). The American National Corpus: More than the Web can Provide. Proceedings of the Third Language Resources and Evaluation Conference, Las Palmas, 839- 844.

Reppen, R. & Ide, N. (in press). The American National Corpus: An Update and Overview of the First Release. Journal of English Linguistics.

Sjöbergh, Jonas (2003). Combining POS-taggers for improved accuracy on Swedish text, NoDaLiDa 2003, Reykjavik.

Tufis, Dan (2000). Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. Proceedings of the Second International Conference on Language Resources and Evaluation, 1105-111.